

CHAPTER 17: Tests of Significance: The Basics

**Basic Practice of
Statistics**

7th Edition

Lecture PowerPoint Slides

In chapter 17, we cover ...

- The reasoning of tests of significance
- Stating hypotheses
- P -value and statistical significance
- Tests for a population mean
- Significance from a table*
- Resampling: Significance from a simulation*

Statistical inference

- Confidence intervals are one of the two most common types of statistical inference. Use a confidence interval when your goal is to estimate a population parameter.
- The second common type of inference, called *tests of significance*, has a different goal: to assess the evidence provided by data about some claim concerning a population.

A **test of significance** is a formal procedure for comparing observed data with a claim (also called a hypothesis) whose truth we want to assess.

- Significance tests use an elaborate vocabulary, but the basic idea is simple: *an outcome that would rarely happen if a claim were true is good evidence that the claim is not true.*

The reasoning of tests of significance

- Artificial sweeteners in colas gradually lose their sweetness over time. Manufacturers test for loss of sweetness on a scale of -10 to 10, with negative scores corresponding to a gain in sweetness, positive to loss of sweetness.
- Suppose we know that for any cola, the sweetness loss scores vary from taster to taster according to a Normal distribution with standard deviation $\sigma = 1$. The mean μ for all tasters measures loss of sweetness and is different for different colas.
- Here are the sweetness losses for a cola currently on the market, as measured by 10 trained tasters:

2.0 0.4 0.7 2.0 -0.4 2.2 -1.3 1.2 1.1 2.3

- The average sweetness loss is given by the sample mean $\bar{x} = 1.02$. Most scores were positive. That is, most tasters found a loss of sweetness. But the losses are small, and two tasters (the negative scores) thought the cola gained sweetness. Are these data good evidence that the cola lost sweetness in storage?

The reasoning of tests of significance

- We make a claim and ask if the data give evidence *against* it. We seek evidence that there is a sweetness loss, so the claim we test is that there is *not* a loss. In that case, the mean loss for the population of all trained testers would be $\mu = 0$.
- If the claim that $\mu = 0$ is true, the sampling distribution of \bar{x} from 10 tasters is Normal with mean $\mu = 0$ and standard deviation

$$\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{10}} = 0.316$$

- This is like calculations we did in Chapter 15—we can locate our \bar{x} of 1.02 in this distribution and comment on whether it is surprising.

Stating hypotheses

- A significance test starts with a careful statement of the claims we want to compare.
- The claim tested by a statistical test is called the **null hypothesis** (H_0). The test is designed to assess the strength of the evidence against the null hypothesis. Often the null hypothesis is a statement of “no difference.”
- The claim about the population that we are trying to find evidence for is the **alternative hypothesis** (H_a). The alternative is **one-sided** if it states that a parameter is *larger* or *smaller* than the null hypothesis value. It is **two-sided** if it states that the parameter is *different from* the null value (it could be either smaller or larger).
- In the sweetness example, our hypotheses are
$$H_0 : \mu = 0$$
$$H_a : \mu > 0$$
- The alternative hypothesis is one-sided because we are interested only in whether the cola lost sweetness.

Example

- Does the job satisfaction of assembly workers differ when their work is machine-paced rather than self-paced? Assign workers either to an assembly line moving at a fixed pace or to a self-paced setting. All subjects work in both settings, in random order. This is a matched pairs design. After two weeks in each work setting, the workers take a test of job satisfaction. The response variable is the difference in satisfaction scores, self-paced minus machine-paced.
- The parameter of interest is the mean μ of the differences in scores in the population of all assembly workers. The null hypothesis says that there is no difference between self-paced and machine-paced work, that is,

$$H_0 : \mu = 0$$

- The authors of the study wanted to know if the two work conditions have different levels of job satisfaction. They did not specify the direction of the difference. The alternative hypothesis is therefore *two-sided*:

$$H_a : \mu \neq 0$$

P-value and statistical significance

- The null hypothesis H_0 states the claim that we are seeking evidence against. The probability that measures the strength of the evidence against a null hypothesis is called a **P-value**.
- A **test statistic** calculated from the sample data measures how far the data diverge from what we would expect if the null hypothesis H_0 were true. Large values of the statistic show that the data are not consistent with H_0 .
- The probability, computed assuming H_0 is true, that the statistic would take a value as extreme as or more extreme than the one actually observed is called the **P-value** of the test. The smaller the P-value, the stronger the evidence against H_0 provided by the data.
- Small P-values are evidence against H_0 because they say that the observed result is unlikely to occur when H_0 is true.
- Large P-values fail to give convincing evidence against H_0 because they say that the observed result could have occurred by chance if H_0 were true.

P -value and statistical significance

- Tests of significance assess the evidence against H_0 . If the evidence is strong, we can confidently reject H_0 in favor of the alternative.
- Our conclusion in a significance test comes down to:
 - P -value small \rightarrow reject $H_0 \rightarrow$ conclude H_a (in context)
 - P -value large \rightarrow fail to reject $H_0 \rightarrow$ cannot conclude H_a (in context)
- There is no rule for how small a P -value we should require in order to reject H_0 — it's a matter of judgment and depends on the specific circumstances. But we can compare the P -value with a fixed value that we regard as decisive, called the significance level. We write it as α , the Greek letter alpha. When our P -value is less than the chosen α , we say that the result is statistically significant.
- If the P -value is smaller than alpha, we say that the data are **statistically significant at level α** . The quantity α is called the **significance level** or the **level of significance**.

Tests of significance

TESTS OF SIGNIFICANCE: THE FOUR-STEP PROCESS

- **STATE:** What is the practical question that requires a statistical test?
- **PLAN:** Identify the parameter, state null and alternative hypotheses, and choose the type of test that fits your situation.
- **SOLVE:** Carry out the test in three phases:
 1. **Check the conditions** for the test you plan to use.
 2. Calculate the **test statistic**.
 3. Find the ***P*-value**.
- **CONCLUDE:** Return to the practical question to describe your results in this setting.

Tests for a population mean

Z TEST FOR A POPULATION MEAN

- Draw an SRS of size n from a Normal population that has unknown mean μ and known standard deviation σ . To test the null hypothesis that μ has a specified value,

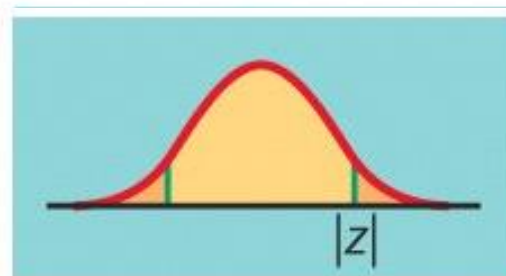
$$H_0: \mu = \mu_0$$

- Compute the one-sample z test statistic.

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- In terms of a variable Z having the standard Normal distribution, the approximate P -value for a test of H_0 against

- $H_a: \mu > \mu_0$ is $P(Z \geq z)$
- $H_a: \mu < \mu_0$ is $P(Z \leq z)$
- $H_a: \mu \neq \mu_0$ is $2 \times P(Z \geq |z|)$



Example

Executives' blood pressures

- **STATE:** NCHS reports that the systolic blood pressure for males 35 to 44 years of age has mean 128 and standard deviation 15. The medical director of a company finds, for the 72 executives in this age group, that

$$\bar{x} = 126.07$$

Is this evidence that the company's executives have a different mean systolic blood pressure from the general population?

- **PLAN:** Take μ to be the mean among all executives. We want to test the hypotheses

$$H_0: \mu = 128$$

$$H_a: \mu \neq 128$$

The alternative is two-sided because the medical director did not have a particular direction in mind before examining the data.

Example, cont'd.

Executives' blood pressures, cont'd.

- **SOLVE:** As part of the “simple conditions,” suppose we are willing to assume that executives' systolic blood pressures follow a Normal distribution with standard deviation $\sigma = 15$. Software can now calculate z and P for you. Going ahead by hand, the **test statistic** is

$$\begin{aligned} z &= \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \\ &= \frac{126.07 - 128}{15 / \sqrt{72}} = -1.09 \end{aligned}$$

Using Table A or software, we find that the P -value is 0.2758.

- **CONCLUDE:** More than 27% of the time, an SRS of size 72 from the general male population would have a mean systolic blood pressure at least as far from 128 as that of the executive sample. The observed $\bar{x} = 126.07$ is therefore not good evidence that executives differ from other men.

Significance from a table*

- Statistics in practice uses technology to get P -values quickly and accurately. In the absence of suitable technology, you can get approximate P -values by comparing your test statistic with critical values from a table.

SIGNIFICANCE FROM A TABLE OF CRITICAL VALUES

- To find the approximate P -value for any z statistic, compare z (ignoring its sign) with the critical values z^* at the bottom of Table C. If z falls between two values of z^* , the P -value falls between the two corresponding values of P in the “One-sided P ” or the “Two-sided P ” row of Table C.

Resampling: Significance from a simulation*

- We saw in Section 15.3 that we can approximate the sampling distribution of \bar{x} by taking a very large number of SRS's of size n and constructing the histogram of the values of the sample means, \bar{x} .
- A corresponding method of taking a large number of repeated SRS's from the population distribution when the null hypothesis is true and using these to approximate P -values is sometimes referred to as **resampling**.
- All we need to know is the population distribution under the assumption that the null hypothesis is true. We then resample, using software, many times from this population distribution, compute the value of the sample statistic for each sample, and determine the proportion of times we obtained sample values *as or more extreme* than that of our actual data. This proportion is an estimate of the P -value.

Resampling: Significance from a simulation*

- Comments about resampling:
 - First, we must resample in the same manner that we obtain our data. If our data are obtained by an SRS, we resample by taking repeated SRS's from the population distribution determined by the null hypothesis.
 - Second, resampling only provided an estimate of a P -value. Repeat the resampling and you will obtain a different estimate. Accuracy of the estimate is improved by taking a larger number of samples to estimate the sampling distribution.
 - Finally, resampling requires the use of software.